第46卷 第3期
2023年3月
干旱区地理
ARID LAND GEOGRAPHY
Vol. 46 No. 3
Mar. 2023

干旱地区"代表性人口格网数据集"精度研究
——以甘宁青地区为例

肖东升[1,2]， 王 宁[1,2]， 刘志成[1,2]

（1. 西南石油大学土木工程与测绘学院，四川 成都 610500；
2. 西南石油大学测绘遥感地理信息防灾应急研究中心，四川 成都 610500）

摘 要：高精度的人口格网数据集在风险评价、灾害应急、生态环境保护、区域发展与规划等领域具有重要价值。输入数据精度和模型选择的不同导致其具有不同的特点与优势，因此评价代表性数据集的精度，分析数据集的适用条件意义重大。研究评估了世界人口（WorldPop）数据集和世界第四版网格化人口（GPWv4）数据集在中国西北干旱地区甘肃省、宁夏回族自治区和青海省的精度；以中国人口普查数据的最佳可用单位（乡镇行政区划）为研究单元，将WorldPop和GPWv4数据集与2020年第七次人口普查数据进行相关性分析，计算统计误差和相对误差的空间分布，定量地评价各个数据集的精度；通过目视估计定性地分析数据集的映射性能，最后讨论了数据集的误差来源。统计误差结果表明：WorldPop数据集的精度更高，其相关系数（r）、均方根误（RMSE）、平均绝对误差（MAE）和平均绝对百分比误差（MAPE）分别达到0.76、23016、0.73和0.60，而GPWv4数据集的上述统计结果分别为0.70、22297、0.75和0.58。同时，由相对误差的空间分布可知，WorldPop数据集准确估计的区域更多。目视估计结果表明：2种人口格网数据集的映射性能类似，都具有东部人口稠密、西部人口稀疏的特点。针对干旱地区人口格网数据集精度的评价研究，有利于分析数据集的误差来源，指导数据集的合理使用。在未来研究中，使用人类生活的辅助数据，生成干旱地区特有的人口分布模式，从而提高西北干旱区域人口数据集的精度。

关 键 词：人口格网数据集；GPWv4数据集；WorldPop数据集；精度评价；中国西北干旱地区

文章编号：1000－6060(2023)03－0505－10(0505～0514)

人口数据是风险评价、灾害应急、生态环境保护、区域发展与规划等领域的重要基础数据，也是人口空间化研究的主要数据源[1-2]。传统上人口数据的获取通常是以行政区为基本单元收集的全国人口普查数据，具有空间分辨率低，时间分辨率低的特点，且无法充分揭示行政区内人口数据的空间异质性[3]。同时，以行政区为基本单元的统计人口数据无法与以格网等基础地理单元数据耦合，难以满足空间分析、统计的需求[4-5]。因此，建立能反映真实人口空间分布的人口格网数据集，预测人口数据及其时空分布，具有重要的理论与现实意义。目前，已有研究者建立了众多的人口格网数据集，其中仍被广泛使用的人口格网数据集包含LandScan[6]、中国1 km网格人口（CnPop）[7]、全球资源信息数据库（UNEP/GRID）[8]、全球城乡测绘项目（GRUMP）[9]、世界第四版网格化人口数据集（GPWv4）[10]、OpenPopGrid[11]、全球人类居住区（GHS）[12]和世界人口（WorldPop）数据集[13]等，这些数据集被灵活的利用在各类研究中。当前仅有WorldPop和GPWv4数据集中数据更新到2020年。

人群空间分布的复杂性以及生成数据集的模型的局限性，导致了人口格网数据集与实际人口必

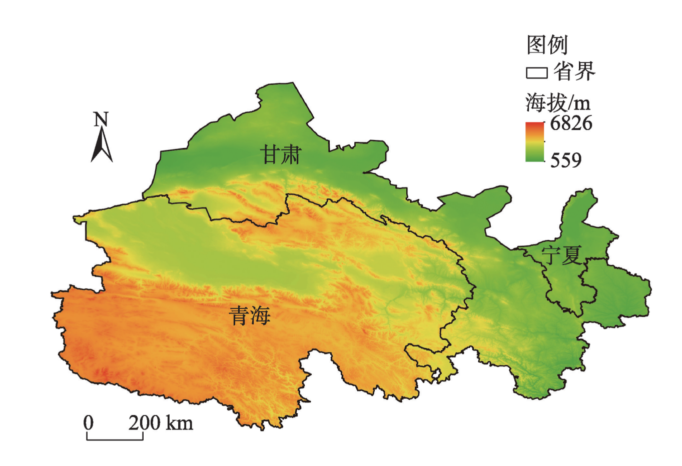
收稿日期：2022-07-14； **修订日期**：2022-09-07
基金项目：国家自然科学基金项目（51774250）；四川省科技计划项目（2019JDR0112）；西南石油大学测绘遥感地信与防灾应急青年科技创新团队（2019CXTD07）；四川省科技创新苗子工程项目（2020046）；四川省科技厅区域创新合作项目（23QYCX0053）资助
作者简介：肖东升（1974-），男，博士，教授，主要从事地震压埋人员定位、人口空间化与防灾减灾技术研究. E-mail: xiaodsxds@163.com
通讯作者：王宁（1997-），男，硕士研究生，主要从事地理信息系统、区域人口空间化与防灾减灾技术研究. E-mail: 2978404858@qq.com

然存在误差，评估现有人口格网数据集的精度，不仅有利于数据生产者发掘数据集的短板，也有利于数据使用者了解数据集的特性，针对研究需要选择研究数据[14-15]。评价人口格网数据集的精度仍然是一项具有挑战性的工作，针对解决这类项目主流有2种理论：一种是对产生数据集的模型和数据集进行精度评估[16]。但是，由于人群分布的复杂性与流动性，无法准确地获得格网单元的人口值，所以难度较大[17]。另外一种方法是将人口格网数据统计到行政区内再与人口普查数据进行比较[18-19]。目前，国内外研究者对人口格网数据集的精度已经进行了一些研究，Tatem等[20]通过构建恶性疟原虫疟疾流行性全球地图评价了GRUMP、LandScan、UNEP和GPWv3数据集的准确性，结果显示现有数据集的估计人口分布存在很大差异。王雪梅等[21]以中国黑河流域为研究区，在流域尺度上把GPWv3、UNEP、LandScan和CnPop数据集的人口估计结果与统计数据进行比较分析，结果表明，CnPop数据集精度最高。Bai等[17]使用中国2000年第五次人口普查数据评估了GPWv3、GRUMP、WorldPop和CnPop数据集在全国范围的精度，4个数据集在中国西北地区精度均偏低。Xu等[22]评估了2015年中国西南云南省、广西壮族自治区、贵州省地区GPWv4、GHS、LandScan和WorldPop数据集的精度，同时利用谷歌地球高分辨率图像定性分析了行政区内人口分布。林丹淳等[23]以2010年广东省为例对代表性人口空间分布数据集的精度做出了评价，比较WorldPop、GPWv4数据集和2种中国公里网格人口分布数据集空间分布的一致性。上述研究验证了人口格网数据集在人口密集区具有良好的精度，但在中国西北干旱地区精度大大降低，研究西北干旱地区的人口格网精度以及精度的影响因子，是我们需要考虑的研究方向。此外，之前的研究大多数是基于2000年和2010年的人口数据研究，缺少最新的研究成果。

水资源等环境因素对人群空间分布影响巨大，而中国西北地区位于干旱地区，这导致了西北地区人群聚集模式不同于沿海等气候适宜地区，在以往的研究中干旱地区人口格网数据集的精度都属于未被准确估计区域[17]，所以探究此类地区误差产生的因素与数据集的缺陷，有利于数据开发者提高数据集精度。因此，本文利用2020年第七次人口普查数据为人口真值，对2020年的WorldPop和GPWv4数据集进行精度评价，分析人口格网数据集在中国西北干旱区域的准确性和特征，以弥补研究空白。

1 研究区概况

甘肃省、宁夏回族自治区和青海省(简称甘宁青)位于中国西北干旱地区，北部同内蒙古自治区相接，西北部与新疆维吾尔自治区相邻，西南部与西藏自治区毗连，南部和东南部与四川省接壤，东部与陕西省相连。研究区有27个市级行政区，包括19个地级市和8个民族自治州，共计153个县级行政区，总面积约1.21×10^{6} km^{2}(图1)。甘宁青地区是“丝绸之路”经济带的核心地区，是衔接中国、中亚和欧洲大陆的重要枢纽。该地区自然资源丰富，是中国“西部大开发”战略的重点地区，是“西气东输”“北煤南运”等战略的能源生产外输基地。尽管甘宁青地区拥有广阔的地域与丰富的资源，却是中国人口分布极不协调的地区之一。根据2020年人口普查数据显示，该地区总人口约3.81×10^{7}人，平均人口密度约为31.40人·km^{-2}，远低于中国平均人口密度[24]。因此，研究甘宁青地区人口格网数据集的精度，可以为该地区人口分布和人口空间化研究提供技术支持。



注：该图基于国家测绘地理信息局标准地图服务网站下载的审图号为GS(2019)1822号的标准地图制作，底图边界无修改。下同。

图1 研究区示意图

Fig. 1 Schematic diagram of the study area

2 数据与方法

2.1 人口普查数据

行政区划是国家为方便行政管理而划分等级的区域，行政区划又称行政区域。中国的行政区划

分为省、地、县、乡4级行政区域[25]。本文评价了乡镇尺度人口数据集的精度，这是中国人口普查数据的最佳可用单位。乡级行政边界数据来源于国家基础地理信息中心，比例为1:100000。由于行政区划的调整，一些城镇的行政边界发生了变化。通过对比2020年的行政区划，在ArcGIS软件中对不一致的行政区划进行修改，最终得到2097个乡镇行政区划。乡镇人口普查数据来源于各区、县2020年人口普查公报，可以在行政区域官方网站找到，例如青海省西宁市城中区各个街道乡镇人口普查数据，可以从西宁市城中区人民政府官网处查询（http://www.xncz.gov.cn/info/2953/121813.htm）。2097个乡镇行政区划和乡镇级别的人口普查数据见图2。

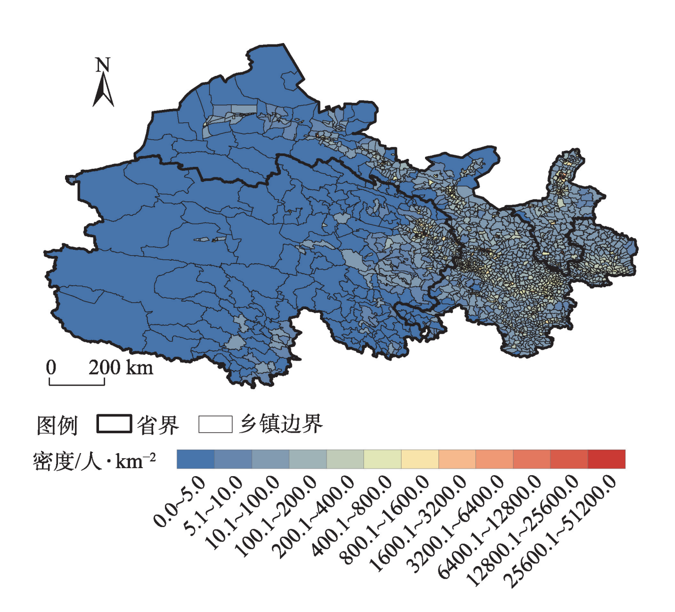


图2　乡镇统计人口数量

Fig. 2　Number of statistical populations in townships

2.2　人口格网数据集

本文选择GPWv4和WorldPop数据集来评估它们在估计人口信息方面的性能。表1给出了4个网格人口分布数据的基本特征。人口网格数据集和乡镇普查数据采用WGS84地心坐标系和Albers等面积投影作为地理参考。

GPWv4数据集是由NASA的社会经济数据和应用中心SEDAC（Socioeconomic Data and Applications Center）发布的全球人口格网数据集。GPWv4数据集在30″（赤道上约1 km）网格单元上模拟了2000、2005、2010、2015年和2020年的全球人口分布。GPWv4数据集的2个基本输入数据是非空间人口数据和空间范围明确的行政边界数据。估计人口是通过人口普查和行政单位的面积比例分配到网格中，同时利用水域作为掩膜，以防止湖泊、河流和冰雪覆盖地区干扰实际的人口分布。

经联合国调整后的GPWv4（A-GPWv4）数据集是由联合国根据联合国人口机构提供的人口数据对原始GPWv4数据集进行调整得到的。

WorldPop数据集的空间分辨率为3″（赤道约为100 m），并提供2000—2020年的年度人口数据估计。它使用例如居住区、夜间卫星图像、道路、植被、地形和土地使用等空间辅助数据集进行建模，以纠正住宅和建成区的分布。然后基于随机森林回归树生成预测加权图层，将官方普查数据重新分布到网格中，实现人口空间化。

经联合国调整后的WorldPop（A-WorldPop）数据集是由联合国根据联合国人口机构提供的人口数据对原始WorldPop数据集进行调整得到的。

2.3　精度评价方法

根据2097个乡镇的行政区划，利用ArcGIS软件的分区统计功能，统计了4个人口格网数据集的估计人口密度。然后利用估计数据和统计数据计算数据集的相对误差（RE），并评价各个人口格网数据集的准确性，RE计算公式如下：

$$RE = \frac{p_g - p_t}{p_t} \qquad (1)$$

表1　人口格网数据集的基本信息

Tab. 1　Basic information of the population grid data sets

数据集	统计指标	时间分辨率	辅助数据	空间分辨率/m	空间化方法
WorldPop	人口密度	2000、2005、2010、2015、2020年	土地使用、居住区、夜间卫星图像、道路、制备、地形	100	随机森林模型
A-WorldPop	联合国调整后的人口密度	2000、2005、2010、2015、2020年	土地使用、居住区、夜间卫星图像、道路、制备、地形、联合国历史人口估计数	100	随机森林模型
GPWv4	人口密度	2000—2020年	水域、行政区划	1000	面积权重法
A-GPWv4	联合国调整后的人口密度	2000—2020年	水域、行政区划、联合国历史人口估计数	1000	面积权重法

为了评估人口网格数据集的准确性,计算估计值和统计值的相关系数(r)、均方根误(RMSE)、平均绝对误差(MAE)和平均绝对百分比误差(MAPE)[17],计算公式如下:

$$r=\frac{\sum_{i=1}^{n}(p_g-\overline{p_g})(p_t-\overline{p_t})}{\sqrt{\sum_{i=1}^{n}(p_g-\overline{p_g})^2\sum_{i=1}^{n}(p_t-\overline{p_t})^2}} \tag{2}$$

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(p_g-p_t)^2}{n}} \tag{3}$$

$$\mathrm{MAE}=\frac{\sum_{i=1}^{n}\left|p_g-p_t\right|/p_t}{n} \tag{4}$$

$$\mathrm{MAPE}=\frac{\sum_{i=1}^{n}(p_g-p_t)/p_t}{n} \tag{5}$$

式中:p_g 为每个格网数据集乡镇行政区域内的估计人口;p_t 为对应乡镇的普查人口数据;n 为行政区划的数量。

3 结果与分析

3.1 目视估计

图3分别展示了研究区内GPWv4、WorldPop、A-GPWv4和A-WorldPop 4种人口格网数据集的人口分布情况。为了在同一分辨率下对比分析4种数据集,在ArcGIS中预处理WorldPop数据集,使其分辨率为1 km。与统计人口数据相比,图3中的人口网格密度数据也呈现出类似的趋势:东部人口分布较为密集,而西部人口分布较为稀疏。受气候和地理环境的影响,西北地区人口分布呈现出以大城市为中心的放射性分布特征,例如宁夏回族自治区的银川市、吴忠市和中卫市,甘肃省的兰州市、天水市、武威市、张掖市、酒泉市,青海省的西宁市和格尔木市是主要的人口聚集地。而与东部人口高度集中分布相比,西部地区面积大,人口稀疏,人口分布相对分散,大多数地区人口密度低于5人·km^{-2}。

在映射性能的比较中,更改分辨率后的WorldPop和GPWv4数据集的映射视觉效果与普查人口数据映射视觉效果类似,说明预测人口格网数据集大体上反映了实际人口分布,具有良好的精确度。而GPWv4和WorldPop数据集的视觉效果进行比较时也有显著差异。GPWv4数据集是根据现有的人口普查数据和每个行政单位的人口增长率计算出的一个行政单位的人口估计,利用面积权重法去估计格网人口。这种简单的人口数据区域加权分配和较低的分辨率导致了GPWv4数据集有较明显的拼接

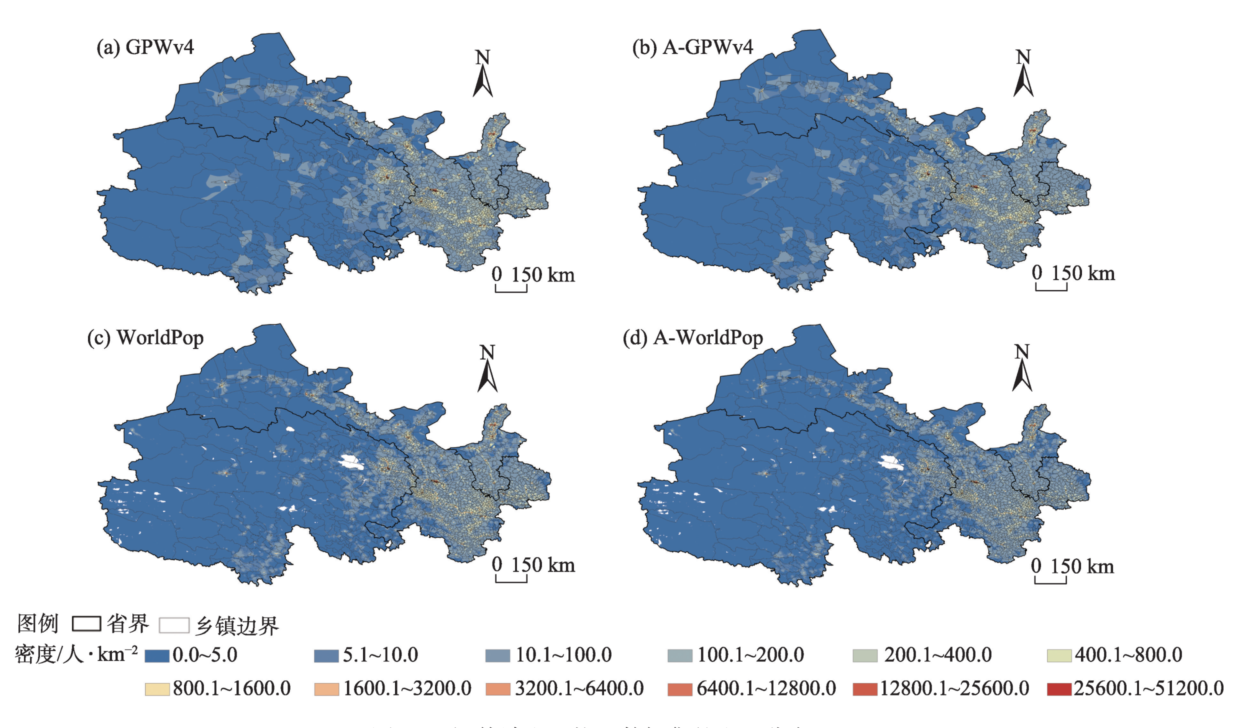


图3 目视估计人口格网数据集的人口分布

Fig. 3 Visually estimated population distributions of population grid data sets

感，WorldPop数据集的空间连续性更好。与GPWv4数据集相比，WorldPop数据集提供了更详细的空间异质性。将广泛可用的遥感和地理空间数据集（如居民点位置、土地覆盖、建筑地图、植被）结合起来，形成模型的对称权重，然后使用随机森林模型生成大约100 m空间分辨率的人口密度网格，这也导致了WorldPop数据集的中心集聚形态更加明显。同时，由于辅助数据集的粗分类作用，人口密度在WorldPop数据集上的变化并不平稳。总而言之，无论是人口稠密的东部地区还是人口稀疏的西部地区，WorldPop数据集对人口分布差异的描述总是优于GPWv4数据集。

3.2 统计分析

表2显示了GPWv4、WorldPop、A-GPWv4和A-WorldPop数据集的乡镇人口密度误差统计结果。A-WorldPop数据集的r最高（0.75），其次是WorldPop、A-GPWv4和GPWv4数据集，对应r分别为0.74、0.69和0.64。A-WorldPop数据集的RMSE最小（16164），A-GPWv4、WorldPop和GPWv4数据集分别为22506、23654和26598。WorldPop数据集的MAE和MAPE

表2 4种人口格网数据集的误差统计

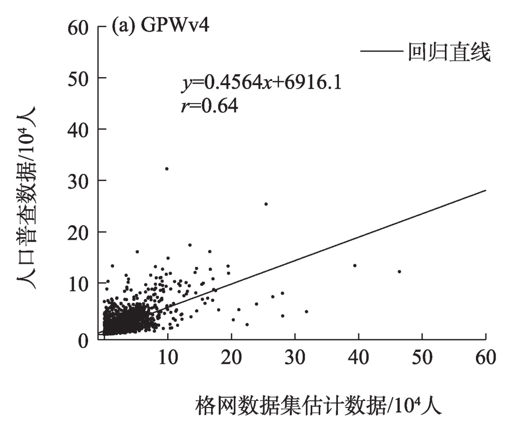
Tab. 2 Error statistics of the four population grid data sets

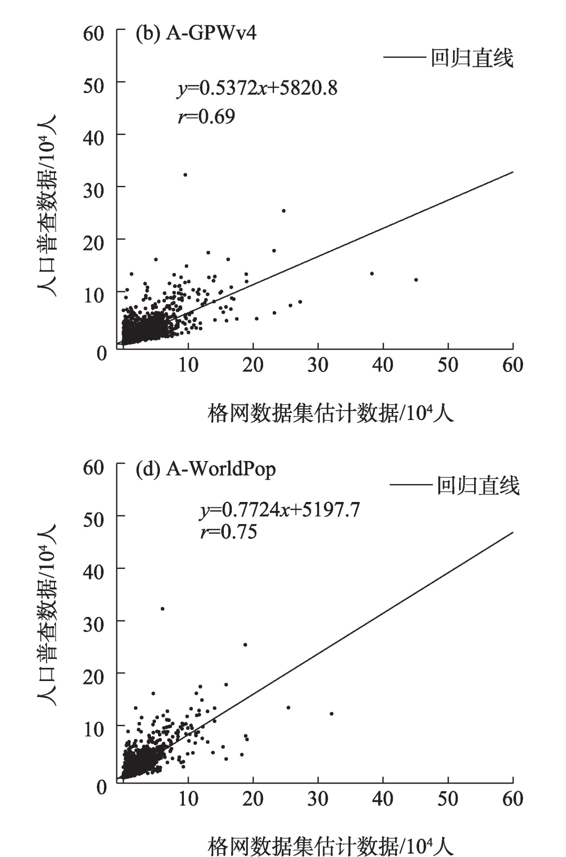
数据集	r	RMSE	MAE	MAPE
GPWv4	0.64	26598	1.41	1.26
A-GPWv4	0.69	22506	0.81	0.64
WorldPop	0.74	23654	0.81	0.69
A-WorldPop	0.75	16164	0.45	0.16

注：r为相关系数；RMSE为方根误差；MAE为平均绝对误差；MAPE为平均绝对百分比误差。下同。

（0.81和0.69）远低于GPWv4数据集（1.41和1.26），而A-WorldPop数据集（0.45和0.16）和A-GPWv4数据集（0.81和0.64）经联合国调整后均有所改善。总体而言，WorldPop数据集在人口网格数据方面优于GPWv4数据集，而联合国调整后WorldPop和GPWv4数据集的精度有所提高。值得注意的是，这4种数据集的RMSE值都较大，表明人口密度估计误差高度离散；而4种数据集的MAE值都很小，表明数据集的总体精度高。

图4显示了4种人口网格数据集和2020年人口普查数据的人口密度散点图，从4个散点图来看，总





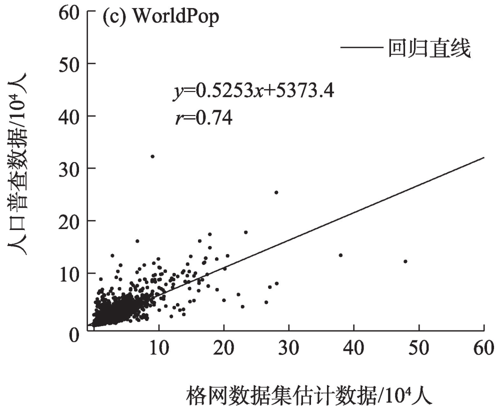


图4 乡镇级别的统计人口与格网数据集的预测人口的相关性分析

Fig. 4 Correlation analysis between statistical population at township level and predicted population in grid datasets

体趋势线接近1:1,说明人口网格数据集与人口普查数据具有良好的一致性。其中,A-WorldPop数据集统计样本更接近1:1,其r也是最大的(0.75),而GPWv4数据集样本离散性最强,而r也是最小的(0.64)。

4种数据集中有一些样本误差较大的异常值,这可能是由于非自然因素导致的人口迁移造成的,例如泥石流等自然灾害或居民点拆迁。同时,乡镇的人口增长率与网格数据的人口增长率之间的差异也会导致预测人口数据的偏差。研究发现每个网格数据集中有异常值的区域虽然不同,但它们具有一系列相似的特征,这可能是由于数据集的生成方式不同以及输入变量的差异所致。为了保证实验的准确性,将每个数据集中误差最大的1%的数据(每个数据集中21个样本)去除,并重新计算4种数据集的误差统计量。

如表3所示,在去除异常值后,4种网格数据集的精度都有所提高,尤其是GPWv4数据集有了较大的提高,而WorldPop数据集的r变化不大。与此同时,网格数据集的RMSE有所提高,MAE和MAPE变化明显。总体表现最好的是A-WorldPop和A-GPWv4数据集,其次是WorldPop和GPWv4数据集。

如图5所示,4种数据集的相对误差空间分布呈现出相似的趋势:严重高估区域在西部占主导地

表3 4个人口格网数据集去除1%的异常值后的误差统计

Tab. 3 Error statistics of four population grid data sets after removing 1% outliers

数据集	r	RMSE	MAE	MAPE
GPWv4	0.70	22297	0.75	0.58
A-GPWv4	0.75	15798	0.40	0.10
WorldPop	0.76	23016	0.73	0.60
A-WorldPop	0.76	15798	0.40	0.10

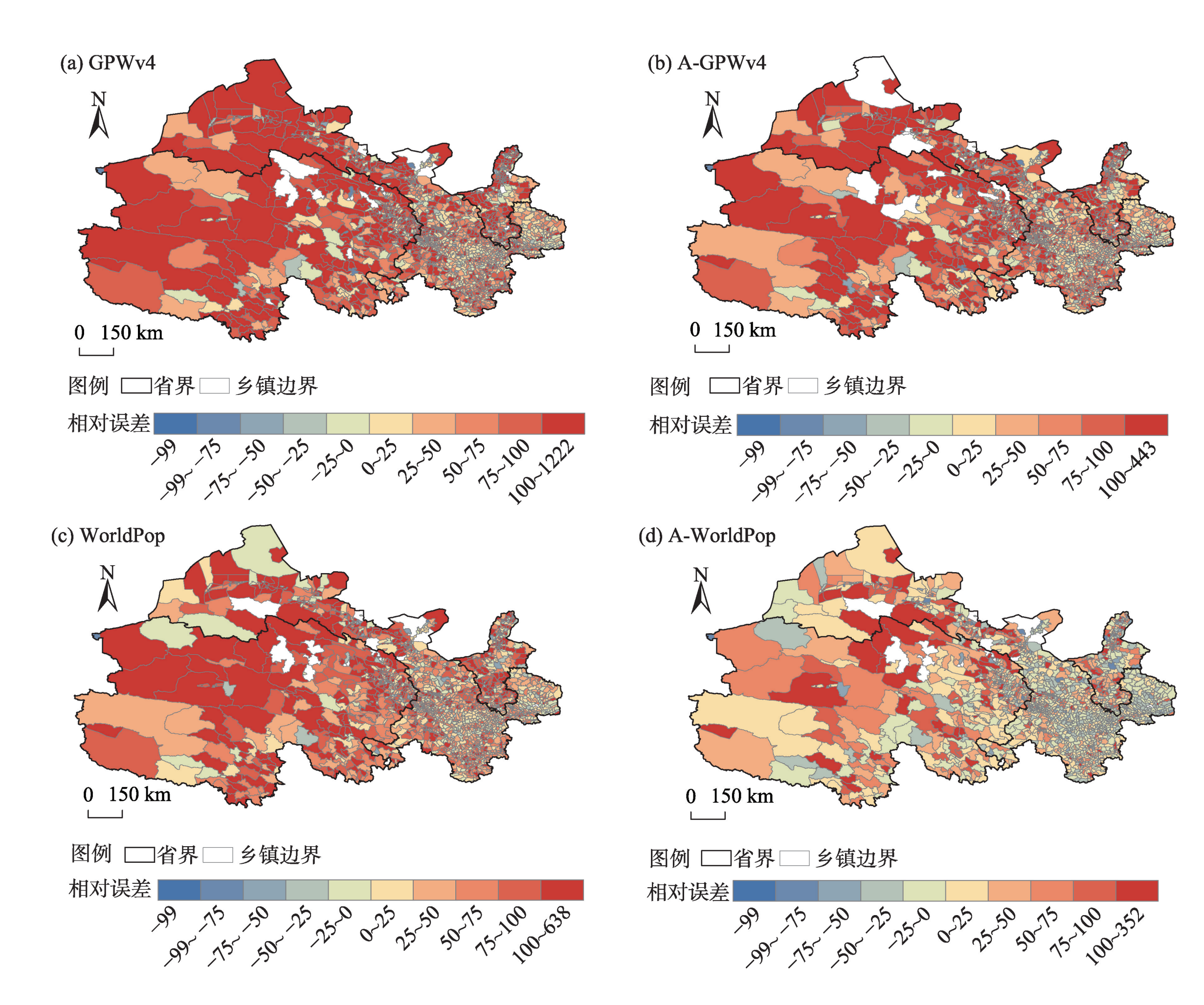


图5 4个数据集相对误差的空间分布

Fig. 5 Spatial distributions of relative errors of the four data sets

位，而东部分布相对均衡。相对误差在-25~25之间被认为是准确估计；相对误差在-50~-25之间及在25~50之间分别被认为是低估及高估；相对误差在-100~-50和50~100之间分别被认为是严重低估和严重高估。从图5a可以看出，GPWv4数据集中严重高估区域明显多于其他数据集。特别是西部和北部面积较大的乡镇行政单位大多属于严重高估区域，而很少有区域被低估。高估区域主要分布在丘陵、戈壁和高原地区，而准确估计则集中在平原等城市密集地区。这表明在平原区域GPWv4数据集采用的面积加权法具有较好的精度。图5b和图5c分别为A-GPWv4和WorldPop数据集的相对误差分布。可以看出，2组数据集的整体性能相似，高估区域分布大致相同，但WorldPop数据集的准确估计区域要大于A-GPWv4数据集。从图5d可以看出，准确估计面积成为主体，高估面积大大减少，但低估面积增加。横向比较GPWv4和WorldPop数据集的结果表明，联合国调整后的网格数据集表现较好，对乡镇级别行政单元的高估减少。特别是对于WorldPop数据集来说，经调整后的数据集准确估计区域占主导地位。纵向比较GPWv4和WorldPop数据集，发现WorldPop数据集总体精度较高，特别是A-WorldPop数据集中精确估计区域的比例远高于高估区域和低估区域。

为了更直观地显示数据集的误差分布，绘制了4种数据集的泰勒图，并且统计了每个数据集的误差分布区间。如图6所示，A-WorldPop数据集精度

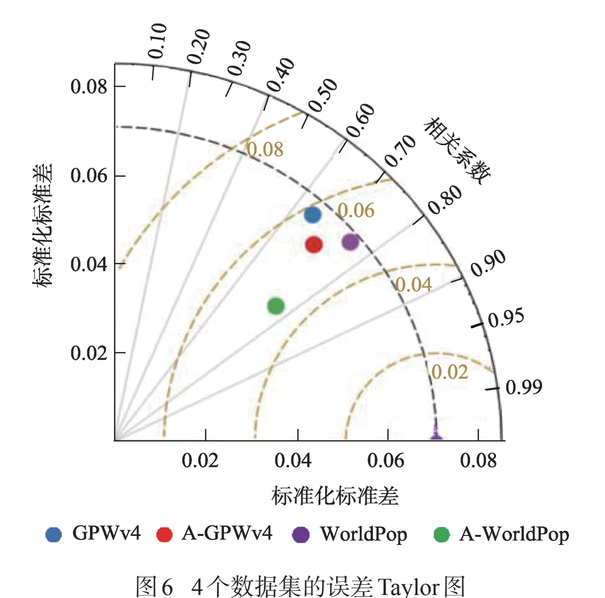


图6　4个数据集的误差Taylor图

Fig. 6　Error Taylor diagram of the four data sets

远超其余3个数据集。图7展示了4种数据集相对误差分布情况，可以清晰的看出，A-WorldPop数据集准确估计区域在4个数据集中占比最大，同时A-WorldPop数据集总体被低估，而GPWv4、A-GPWv4和WorldPop数据集总体被高估。A-WorldPop数据集在准确估计、低估和严重低估区域占比都大于其他3个数据集，而在高估区域与其他3个数据集也将近持平。总体而言，GPWv4、A-GPWv4和WorldPop数据集高估了中国乡镇的总人口，而A-WorldPop数据集更准确。

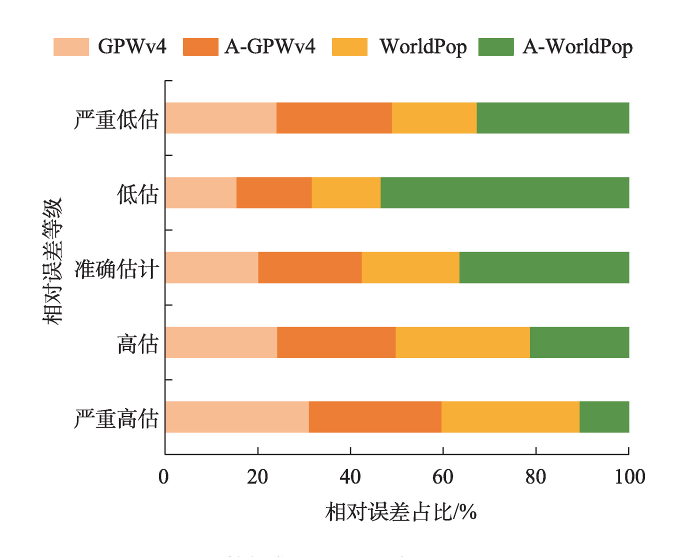


图7　4个数据集相对误差占比堆积柱状图

Fig. 7　Stacked histogram of relative error percentage of the four data sets

4　讨 论

本研究基于人口普查数据研究了中国西北地区甘肃省、宁夏回族自治区和青海省的WorldPop和GPWv4数据集的精度差异。2种人口格网数据集在西北部高估和东部存在低估的现象。在乡镇尺度上，WorldPop数据集在东部人口密集区域表现良好。WorldPop数据集的空间分辨率有100 m与1 km 2种，相比于GPWv4数据集仅有1 km的空间分辨率，提供了更多的选择，同时也提高了精度，能够更为准确地描述人口空间分布且反映出更多细节信息，在高人口密度地区有良好的表现。而GPWv4数据集在西北部人口稀疏区域和中等人口密度区域未能表现出较好的精度，存在严重高估现象。由于其低空间分辨率和面积权重法的限制，GPWv4数据

集无法反映行政区内人群的真实分布情况，但是其特殊的数据生产方式，使其具备了行政区内人口总数更为精确的特点，在低人口密度地区和最小研究单元大于行政区的研究中有良好表现。数据源和生产方式是导致这2种数据集在该区域人口分布格局出现差异的主要原因。

（1）生产方法

WorldPop数据集使用对数转换的人口密度和消除零计数单位有助于随机森林算法在数据中找到良好的分割，同时使人口密度在大多数情况下分布更加均匀。但是这样产生的一个没有零单元格的dasymetric密度加权层会导致低人口密度区域被高估，高人口密度区域被低估。而GPWv4数据集使用简单面积加权法分配乡镇行政区内的人口，即假设各行政单元人口均匀分布，其精度依赖于输入数据的精细[20]。GPWv4数据集适合用于长时序、最小研究单元大于街镇的研究，在沿海地区等人口分布较均匀的城镇具有较高的精度[23]。

（2）数据源

对数据集精度影响的另一重要因素是其使用的数据源。其中，主要输入数据是基于人口普查的人口数据，并且经过联合国适当的调整以符合目标年份的全球总人口估计数。但由于自然和社会经济条件的空间差异，研究区城镇单位的人口增长率与全国平均水平存在差异，研究区城镇单位的实际调整比例也不同于全国平均水平。因此，全国调整会导致估计人口与城镇统计人口之间的差异。此外，中国近期频繁进行行政区划调整，这可能导致人口格网数据集与镇级统计人口数据之间存在额外的差异。同时辅助数据的输入也会影响数据集的精度。此外，WorldPop数据集使用居住区、夜间卫星图像、道路、植被、地形和土地使用等空间辅助数据集进行建模，以纠正住宅和建成区的分布，同时加入的夜间灯光数据也能使人口格网数据集精度提高；GPWv4数据集的2个基本输入数据仅有非空间人口数据和空间范围明确的行政边界数据。

需要注意的是，中国西北的地理环境导致了西北地区人口分布呈现“大聚集、小群居”的格局，并且人口分布与经济文化、历史基础、水资源等生活资源分布有着密切的联系[26]。从根本上来说，准确性评估的估计误差是由中国人口分布的复杂性造成的。我们推测有3个主要原因影响人口分布的准确性。第一是西北地区农田和小面积农村错落分布，无法从土地利用数据中提取到有用信息确定人口分布[27]；第二是西北地区生态环境脆弱，对人类活动反应敏感，因此人口流动性强[28]；第三是行政区划内的人口被假设为一个固定值，无法反映真实人口分布的空间异质性[29]。若要提高西北地区人口格网数据集的精度，这需要引入地理环境、人口聚集模式等影响人类生活的辅助数据改进数据集。

5 结论

本文比较了GPWv4和WorldPop 2种人口格网数据集在人口分布呈现“大聚集、小群居”特殊格局的中国西北干旱地区人口空间化的精确度。主要结论如下：

（1）基于GPWv4和WorldPop 2种人口格网数据集与普查人口数据，对比了两者在研究区的映射性能，2种数据集与人口普查数据视觉效果类似，都具有东部人口稠密、西部人口稀疏的特点。

（2）通过对GPWv4和WorldPop 2种数据集定性与定量的分析，发现在中国西北地区WorldPop数据集整体表现更好，分类特征明显易区分，具有良好的空间连续性，能够反映出高精度的人口真实空间分布。

（3）人口格网数据集的精度主要受数据源与模型影响。WorldPop数据集采用的随机森林回归模型，通过产生一个没有零单元格的dasymetric密度加权层来估计人口，回归模型产生的平均数估计值会导致高密度区域被低估，而低密度区域被高估。GPWv4数据集使用的面积权重法假设各行政单元人口均匀分布，其精度依赖于输入数据的精细。

（4）WorldPop数据集更适用于人口密度中等和高人口密度区域的精细化研究，且能刻画出行政区内部的人口异质性；GPWv4数据集适用于最小研究单元大于乡镇行政区划的研究。

（5）在中国西北干旱地区，人口分布与水资源等生活资源分布有着密切的联系，基于水资源分布信息、地理环境、人口聚集模式等影响人类生活的辅助数据生成干旱地区特有的人口分布模式以提高西北干旱区域人口分布精度是未来研究的方向。

参考文献(References)

[1] Fang J Y, Sun S, Shi P J, et al. Assessment and mapping of potential storm surge impacts on global population and economy[J]. International Journal of Disaster Risk Science, 2014, 5(4): 323-331.

[2] Wu X, Yang J, Zhang H. Analyzing spatial autocorrelation of population distribution in different spatial weights: A case of China[J]. Geomatics World, 2017, 24(2): 32-38.

[3] 高义, 王辉, 王培涛, 等. 基于人口普查与多源夜间灯光数据的海岸带人口空间化分析[J]. 资源科学, 2013, 35(12): 2517-2523. [Gao Yi, Wang Hui, Wang Peitao, et al. Population spatial processing for Chinese coastal zones based on census and multiple night light data[J]. Resources Science, 2013, 35(12): 2517-2523.]

[4] 杨小唤, 王乃斌, 江东, 等. 基于空间分析方法的人口空间分布区划[J]. 地理学报, 2002, 57(增刊): 76-81. [Yang Xiaohuan, Wang Naibin, Jiang Dong, et al. Regionalization of population distribution based on spatial analysis[J]. Acta Geographica Sinica, 2002, 57(Suppl.): 76-81.]

[5] 郭洪旭, 黄莹, 赵黛青. 城市居住人口空间分布的模拟研究——以广州市天河区为例[J]. 热带地理, 2013, 33(1): 81-87. [Guo Hongxu, Huang Ying, Zhao Daiqing. A simulation study on the spatial distribution of urban residents: A case study of Tianhe district, Guangzhou[J]. Tropical Geography, 2013, 33(1): 81-87.]

[6] Bhaduri B, Brigh E, Coleman P, et al. LandScan USA: A high-resolution geospatial and temporal modeling approach for population distribution and dynamics[J]. GeoJournal, 2007, 69: 103-117.

[7] 江东, 杨小唤, 王乃斌, 等. 基于RS、GIS的人口空间分布研究[J]. 地球科学进展, 2002, 17(5): 734-738. [Jiang Dong, Yang Xiaohuan, Wang Naibin, et al. Study on spatial distribution of population based on remote sensing and GIS[J]. Advances in Earth Science, 2002, 17(5): 734-738.]

[8] UNEP. Global resource information database[DB/OL]. [2022-09-24]. http://na.unep.net/siouxfalls/datasets/datalist.php.

[9] Center for International Earth Science Information Net-work. Global Rural-Urban Mapping Project (GRUMP), alpha version: Urban extents[R]. New York: Center for International Earth Science Information Network (CIE-SIN), Columbia University of Chicago Magazine, 2004.

[10] Balk D L, Deichmann U, Yetman G, et al. Determining global population distribution: Methods, applications and data[J]. Advances in Parasitology, 2006, 62: 119-156.

[11] OpenGMS. Open data[DB/OL]. [2022-09-24]. https://geomodeling.njnu.edu.cn/dataItem/5cd547056af4560e7433dd2e.

[12] Nieves J J. Modelling global human settlement to better inform annual population modelling[D]. Southampton, United Kingdom: University of Southampton, 2020.

[13] Linard C, Alegana V A, Noor A M, et al. A high-resolution spatial population database of Somalia for disease risk mapping[J]. International Journal of Health Geographics, 2010, 9(1): 1-13.

[14] 柏中强, 王卷乐, 杨飞. 人口数据空间化研究综述[J]. 地理科学进展, 2013, 32(11): 1692-1702. [Bai Zhongqiang, Wang Juanle, Yang Fei. A summary of the research on population data spatialization[J]. Progress in Geography, 2013, 32(11): 1692-1702.]

[15] 董南, 杨小唤, 蔡红艳. 人口数据空间化研究进展[J]. 地球信息科学学报, 2016, 18(10): 1295-1304. [Dong Nan, Yang Xiaohuan, Cai Hongyan. Research progress and perspective on the spatialization of population data[J]. Journal of Geo-information Science, 2016, 18(10): 1295-1304.]

[16] 林丽洁, 林广发, 颜小霞, 等. 人口统计数据空间化模型综述[J]. 亚热带资源与环境学报, 2010, 5(4): 10-16. [Lin Lijie, Lin Guangfa, Yan Xiaoxia, et al. Spatialization models of census data: A review[J]. Journal of Subtropical Resources and Environment, 2010, 5(4): 10-16.]

[17] Bai Z Q, Wang J L, Wang M M, et al. Accuracy assessment of multi-source gridded population distribution datasets in China[J]. Sustainability, 2018, 10(5): 1363, doi: 10.3390/su10051363.

[18] Hall O, Stroh E, Paya F. From census to grids: Comparing gridded population of the world with Swedish census records[J]. The Open Geography Journal, 2012, 5(1): 1-5.

[19] Stevens F R, Gaughan A E, Linard C, et al. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data[J]. PLoS One, 2015, 10(2): e0107042, doi: 10.1371/journal.pone.0107042.

[20] Tatem A J, Campiz N, Gething P W, et al. The effects of spatial population dataset choice on estimates of population at risk of disease[J]. Population Health Metrics, 2011, 9(1): 1-14.

[21] 王雪梅, 李新, 马明国. 基于遥感和GIS的人口数据空间化研究进展及案例分析[J]. 遥感技术与应用, 2004, 19(5): 320-327. [Wang Xuemei, Li Xin, Ma Mingguo. Research progress and case analysis of population data spatialization based on remote sensing and GIS[J]. Remote Sensing Technology and Application, 2004, 19(5): 320-327.]

[22] Xu Y, Ho H C, Knudby A, et al. Comparative assessment of gridded population data sets for complex topography: A study of southwest China[J]. Population and Environment, 2021, 42(3): 360-378.

[23] 林丹淳, 谭敏, 刘凯, 等. 代表性人口空间分布数据集的精度评价——以2010年广东省为例[J]. 热带地理, 2020, 40(2): 346-356. [Lin Danchun, Tan Min, Liu Kai, et al. Accuracy comparison of four gridded population datasets in Guangdong Province[J]. China Tropical Geography, 2020, 40(2): 346-356.]

[24] 石英, 米瑞华. 陕西省人口空间分异研究[J]. 干旱区地理, 2015, 38(2): 368-376. [Shi Ying, Mi Ruihua. Differentiation of population spatial distribution in Shaanxi Province[J]. Arid Land Geography, 2015, 38(2): 368-376.]

[25] 王丰龙, 刘云刚. 准行政区划的理论框架与研究展望[J]. 地理科学, 2021, 41(7): 1149-1157. [Wang Fenglong, Liu Yungang. Theoretical framework and prospect on quasi-administrative division[J]. Scientia Geographica Sinica, 2021, 41(7): 1149-1157.

[26] 米瑞华, 高向东. 陕西省人口分布影响因素的空间计量分析[J]. 干旱区地理, 2020, 43(2): 491−498. [Mi Ruihua, Gao Xiangdong. Factors influencing population distribution in Shaanxi Province using spatial econometric analysis[J]. Arid Land Geography, 2020, 43(2): 491−498.]
[27] Long H L, Li Y R, Liu Y S, et al. Accelerated restructuring in rural China fueled by 'increasing vs. decreasing balance' land-use policy for dealing with hollowed villages[J]. Land Use Policy, 2012, 29(1): 11−22.
[28] Jiao J Y, Zhang Z G, Bai W J, et al. Assessing the ecological success of restoration by afforestation on the Chinese Loess Plateau [J]. Restoration Ecology, 2012, 20(2): 240−249.
[29] Briggs D J, Gulliver J, Fecht D, et al. Dasymetric modelling of small-area population distribution using land cover and light emissions data[J]. Remote Sensing of Environment, 2007, 108(4): 451−466.

Accuracy of "representative population grid dataset" in arid areas: A case of Gansu-Ningxia-Qinghai region

XIAO Dongsheng[1,2], WANG Ning[1,2], LIU Zhicheng[1,2]
(1. School of Civil Engineering and Surveying, Southwest Petroleum University, Chengdu 610500, Sichuan, China; 2. Disaster Prevention and Emergency Response Research Center of Southwest Petroleum University, Chengdu 610500, Sichuan, China)

Abstract: High-accuracy population grid datasets are of great value in the fields of risk assessment, disaster emergency response, ecological environment protection, and regional development. The characteristics and advantages of the datasets vary because of the different input data accuracy and model selection. Therefore, it is of great significance to evaluate the accuracy of datasets and analyze the applicable conditions of datasets. To this end, this study evaluated the accuracy of the WorldPop and GPWv4 datasets in the arid areas of the Gansu Province, Ningxia Hui Autonomous Region, and the Qinghai Province of northwest China. The accuracy of each dataset was quantitatively evaluated by calculating the spatial distribution of statistical and relative errors. Taking the best available unit of census data of China (township administrative division) as the research unit, the correlation analysis was conducted between the WorldPop and GPWv4 datasets and the seventh census data in 2020. The spatial distribution of statistical and relative errors is obtained through correlation analysis to quantitatively evaluate the accuracy of each dataset. Furthermore, the mapping performance of the dataset was qualitatively analyzed by visual estimation. Finally, the error sources of the dataset are discussed. The statistical error results show that WorldPop has higher accuracy than GPWv4. The correlation coefficient (r), root mean square error, average absolute error, and average absolute percentage error of WorldPop are 0.76, 23016, 0.73, and 0.60, respectively, while those of GPWv4 are 0.70, 22297, 0.75, and 0.58, respectively. Concurrently, according to the spatial distribution of the relative error, WorldPop accurately estimates the population of more areas. The visual estimation results show that the mapping performance of the two population grid datasets is similar, with the characteristics of a dense and sparse population in the east and west of the study area, respectively. This study on the accuracy of population grid datasets in arid areas is conducive to analyzing the error sources of datasets and guiding the rational use of datasets. In future research, it would be a beneficial direction to use the auxiliary data of human life to generate a unique population distribution pattern in an arid area to improve the accuracy of population datasets in the northwest arid area of China.

Key words: population grid dataset; GPWv4 data set; WorldPop data set; accuracy evaluation; arid area in northwest China